Position: **Principal Analyst – GenAI Engineer**

Bengaluru, Karnataka, India

Factspan Overview:

Factspan is a pure play data and analytics services organization. We partner with fortune 500 enterprises to build an analytics center of excellence, generating insights and solutions from raw data to solve business challenges, make strategic recommendations and implement new processes that help them succeed. With offices in Seattle, Washington and Bengaluru, India; we use a global delivery model to service our customers. Our customers include industry leaders from Retail, Financial Services, Hospitality, and technology sectors.

Role Overview

We are looking for a LLM Engineer with strong hands-on expertise in building GenAI applications, especially around Retrieval- Augmented Generation (RAG), vector stores, prompt engineering, and secure deployment of LLMs. As part of the offshore team, you will collaborate closely with onsite leads and U.S. client stakeholders to design and implement intelligent AI solutions at scale.

Key Responsibilities:

- Design and develop LLM-driven applications for use cases such as customer support copilots, merchandising intelligence, document summarization, and internal productivity tools.
- Build scalable and secure RAG pipelines using LangChain, LlamaIndex, or similar frameworks.
- Integrate vector databases like FAISS, Weaviate, Milvus, or Pinecone for semantic search.
- Work with LLM APIs (OpenAI, Azure OpenAI, Mistral, Claude) and handle context engineering, chaining, and tool invocation.
- Develop APIs and backend services using Python, FastAPI, and manage deployment workflows using Docker/Kubernetes.
- Ensure clean logging, rate limiting, and data compliance for enterprise readiness.
- Collaborate with DevOps and data engineers to support CI/CD and monitor LLM pipeline performance.
- Work in agile delivery mode with daily scrums, sprint planning, and collaborative backlog grooming.

Key Skills:

- Strong proficiency in Python and one or more GenAI frameworks (LangChain, LlamaIndex, or Haystack).
- Solid experience working with LLM APIs and prompt engineering.
- Hands-on exposure to vector databases and RAG implementation.
- Understanding of token usage, function calling, and context window constraints in LLMs.
- Familiarity with MLOps/DevOps deployment in cloud or on-prem environments (preferably Azure).
- Ability to work collaboratively with cross-location teams and communicate effectively with onsite stakeholders.

Good to have Skills:

- Experience in building AI copilots or internal productivity tools.
- Familiarity with front-end tools like Streamlit or Gradio.
- Exposure to unstructured document processing or OCR.
- Retail domain knowledge is a plus.

Required Qualifications:

- 5+ years of relevant experience in DS/AIML Space
- Atleast 2-3 years of experience in building and deploying GenAI applications
- Bachelor's or Master's in Computer Science, Data Science, or related fields

If you are passionate about leveraging technology to drive business innovation, possess excellent problem-solving skills, and thrive in a dynamic environment, we encourage you to apply for this exciting opportunity. Join us in shaping the future of data analytics and making a meaningful impact in the industry.

Why Should You Apply?

Grow with Us: Be part of a hyper- growth startup with ample number of opportunities to Learn & Innovate.

People: Join hands with the talented, warm, collaborative team and highly accomplished leadership.

Buoyant Culture: Embark on an exciting journey with a team that innovates solutions everyday, tackles challenges head-on and crafts a vibrant work environment